SPATIAL AUDIO DISPLAYS FOR SPEECH COMMUNICATIONS: A COMPARISON OF FREE FIELD AND VIRTUAL ACOUSTIC ENVIRONMENTS

W. Todd Nelson, Robert S. Bolia, Mark A. Ericson, and Richard L. McKinley
Air Force Research Laboratory, Wright-Patterson AFB, OH 45433

The ability of listeners to detect, identify, and monitor multiple simultaneous speech signals was measured in free field and virtual acoustic environments. Factorial combinations of four variables, including audio condition, spatial condition, the number of speech signals, and the sex of the talker were employed using a within-subjects design. Participants were required to detect the presentation of a critical speech signal among a background of non-signal speech events. Results indicated that spatial separation increased the percentage of correctly identified critical speech signals as the number of competing messages increased. These outcomes are discussed in the context of designing binaural speech displays to enhance speech communication in aviation environments.

INTRODUCTION

Numerous researchers (Begault & Wenzel, 1993; McKinley, Ericson, & D'Angelo, 1994; Ricard and Meirs, 1994) have suggested that virtual spatial audio displays may be effective for enhancing an operator's capacity to monitor multiple channels of simultaneous speech. This notion is based on the fact that the spatial separation of acoustic signals improves the intelligibility of speech in noise and assists in the segregation of multiple sound streams, a phenomenon known as the "cocktail party" effect (Cherry, 1953; Wenzel, 1992; Yost, Dye, & Sheft, 1996). As noted by Yost and his colleagues (1996), spatial hearing plays an important role in tasks that characterize the "cocktail party" problem, especially when more than two speech signals are presented simultaneously. Empirical support for this position has been provided by Begault and Wenzel (1993), Crispien and Ehrenberg (1995), Ericson and McKinley (1997), Ricard and Meirs (1994), and Yost et al. (1996). Potential applications for "spatialized" speech displays include communication in air traffic control and tactical environments, traffic collision and avoidance systems, and speech-based navigation and warning systems.

Yet the limits of this putative spatial effect are uncertain; many questions remain unanswered, including: (a) Is the spatial effect independent of the number of competing messages? (b) Are these effects similar in free field and virtual acoustic environments? (c) If not, does the latter impose limits on the efficacy of the spatial effect? The primary objectives of this research were to assess the effects of spatial audio presentation on a listener's ability to detect and identify the critical speech signals among multiple simultaneous competing messages and to compare these effects in free field and virtual acoustic environments. To date, investigations of this sort have been extremely sparse and have also been limited by the number of simultaneous competing messages; hence, it is anticipated that this research will be of interest to scientists and engineers involved in the design of spatial audio displays.

Method

Participants. Four men and four women, naïve to the purposes of the experiment, served as paid participants. Their ages ranged from 19 to 47 years with a mean of 29 years. All participants had normal hearing and localization acuity.

Experimental Design. Two acoustic environment conditions (free field and virtual) were combined factorially with two spatial conditions (spatially-separated and non spatially-separated), eight simultaneous talker conditions (1,2,3,4,5,6,7, and 8), and the sex of the critical speech signal (male and female) to provide 64 experimental conditions. A within-subjects design was employed. The order of experimental conditions was randomized with the constraint that the two spatial conditions were completed during separate experimental sessions. In addition, practical considerations involving the apparatus made it necessary to *block* on the acoustic environment factor, with the free field condition preceding the virtual condition for all participants.

Apparatus. The experiment was conducted at the Air Force Research Laboratory's Auditory Localization Facility (ALF) - a geodesic sphere of radius 2.3 m housed within an anechoic chamber. The geodesic sphere is outfitted with Bose 4.5-in. Helical Voice Coil full-range drivers at each of its 272 vertices. Spatial locations of the speech signals for the free field and virtual audio conditions were restricted to the horizontal plane and are illustrated in Figure 1.

Spatialization of the speech signals in the virtual audio condition was achieved using two of the Air Force Research Laboratory's four-channel 3-D Auditory Display Generators (3-D ADG) coupled to a Polhemus 3Space position tracker. The 3-D ADG employs digital signal processing techniques to encode spatial information in an audio signal and displays the resulting "spatialized" signal over stereo headphones (Sennheiser HD-560).

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
1999	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Spatial Audio Displays for Speech Communications: A Comparison of Free Field and Virtual Acoustic Environments	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Air Force Research Laboratory Wright-Patterson AFB, OH 45433**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**4**	

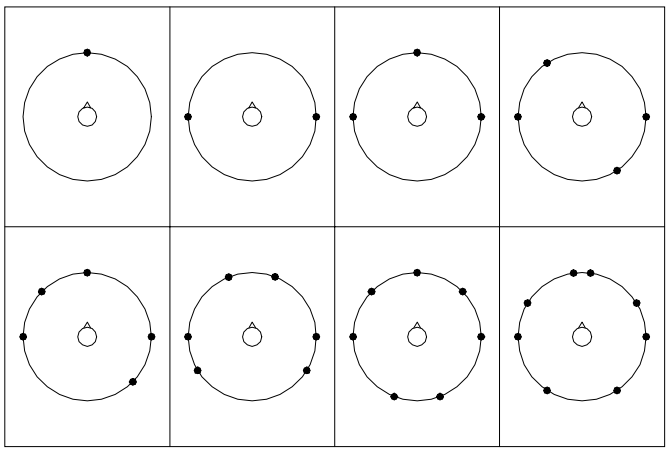


Figure 1. Spatial locations of the speech signals in the free field and virtual audio conditions across the eight talker conditions.

Speech signals were recorded from four male and four female talkers, high-pass filtered at 100 Hz, low-pass filtered at 8 kHz, equated for average power, and edited to ensure synchronous onset. Simultaneous playback of up to eight phrases was achieved using a Tucker-Davis DA3-8 eight-channel digital-to-analog converter.

The spatial locations of the speech signals in the free field and virtual conditions were determined by the locations of the loudspeakers in the ALF, as follows. If, on a given trial, there was only a single talker, the signal always emanated from directly in front of the listener. If there were two or more talkers, the signals were positioned such that the average difference in source-midline distance (SMD) was a maximum for the configuration. If two potential locations had the same average SMD difference, the location was chosen which maximized angular separation. The spatial locations of the speech signals for each of the eight talker conditions are illustrated in Figure 1. For all the non spatially-separated conditions, speech signals originated from 0° azimuth - directly in front of the participant.

Procedure. For each trial, between one and eight speech signals were selected from a set of phrases from a modified version of the Coordinate Response Measure (CRM; Moore, 1981). Each phrase consisted of a call sign (Baron, Ringo, Laker, Charlie, Hopper, Arrow, Tiger, Eagle), a color (Red, White, Green, Blue), and a number (1, 2, 3, 4, 5, 6, 7, 8), embedded within a carrier phrase. Phrases were selected at random with the constraints that 1) the critical signal phrase always contained the call sign "Baron;" and 2) within a given trial on which a critical signal was present, neither talkers nor call signs were repeated. For example, in the 4-talker condition (Figure 2), a listener would hear four different talkers, each uttering a different call sign.

During each trial, participants monitored the simultaneous presentation of multiple speech signals. Their task was to listen for the occurrence of a critical call sign - "Baron" - and to identify the color-number combination that appeared to emanate from the same spatial location as the critical call sign. Participants issued their responses by pressing a key on a response keyboard that was of the appropriate color and number. In the event that the participants did not detect the presence of the critical call sign, they were instructed to press a key that was marked "no signal." Thus, the appropriate response to "Ready Baron Go To Red Six Now" would have been to press the red key labeled with a number six. Fifty percent of the experimental trials included the critical call sign "Baron." Prior to data collection, participants completed several practice sessions.

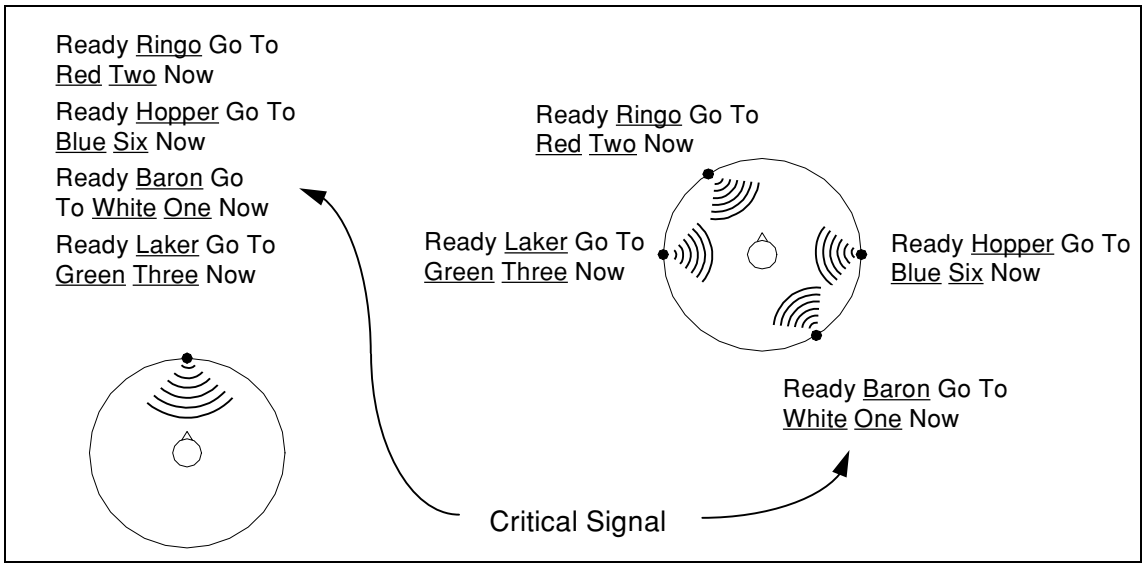


Figure 2. Schematic depicting the non spatially-separated (left) and the spatially separated (right) audio conditions for the 4-talker condition.

Results

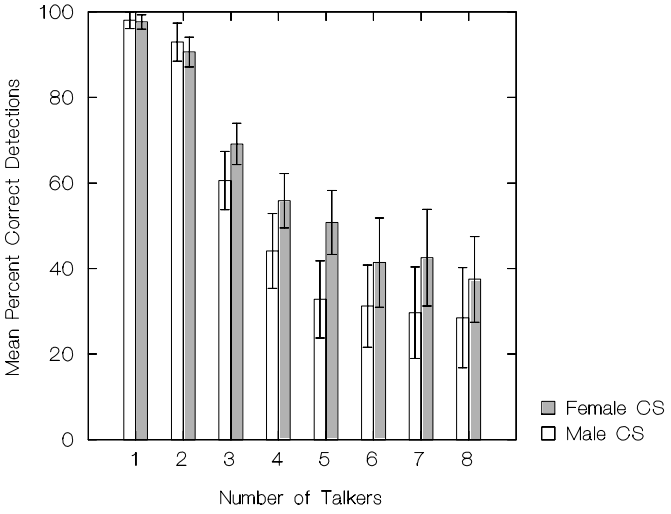
Correct Detections. Detecting the presence of the critical call sign "Baron" when it was present constituted a correct detection. Mean percentage of correct detections were calculated for all experimental conditions and subjected to a 2 (Acoustic Environment) x 2 (Spatial Condition) x 8 (Talker) x 2 (Sex of Critical Signal) repeated measures analysis of variance. The analysis revealed that the main effects of *Talker* and *Sex of Critical Signal* were statistically significant, $F(7,49) = 37.05$, $p<.05$, and $F(1,7) = 9.84$, $p<.05$, respectively. Additionally, the Talker x Sex of Critical Signal interaction was statistically significant, $F(7,49) = 4.44$, $p<.05$. All other sources of variance in the analysis lacked significance ($p>.05$).

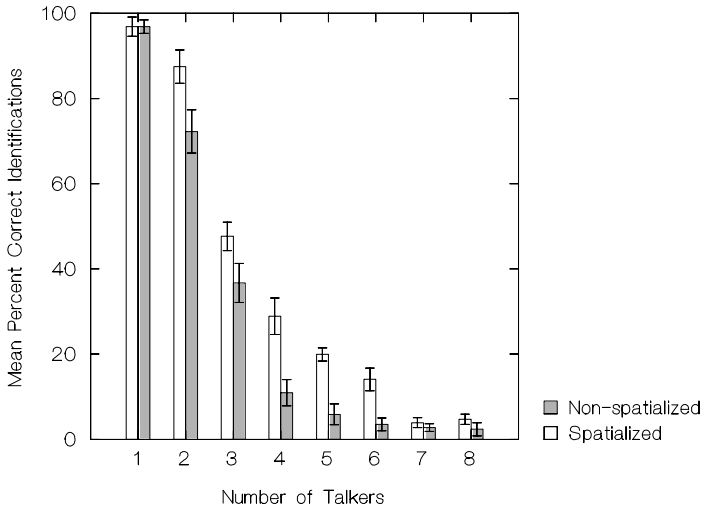
The Talker x Sex of Critical Signal interaction is illustrated in Figure 3, which shows mean percent correct detections plotted for the male and female spoken critical signals within each of the eight talker conditions. It can be observed in the figure that correct detections varied inversely with the number of simultaneous talkers and that female spoken critical signals were detected more frequently than male spoken critical signals when three or more talkers were presented simultaneously. This interpretation was supported by post hoc pairwise comparisons (Bonferroni-adjusted *t*-tests).

Correct Identifications. Mean percentage of correct identifications - i.e., correct detection of the call sign *and* the correct identification of the color and number combination - were calculated for all experimental conditions and analyzed by a similar 2 (Acoustic Environment) x 2 (Spatial Condition) x 8 (Talker) x 2 (Sex of Critical Signal) repeated measures analysis of variance. The analysis revealed significant main effects for Spatial Condition, $F(1,7) = 82.76$, $p<.05$, and Talker, $F(7,49) = 343.75$, $p<.05$, and a significant Spatial Condition x Talker interaction, $F(7,49) = 4.50$, $p<.05$. All other sources of variance in the analysis lacked statistical significance. The Spatial Condition x Talker interaction, which is illustrated in Figure 4, can be explained by noting that spatial separation of the speech signals enhanced performance efficiency when the number or simultaneous talkers was between two and six. Conversely, no advantage for spatial separation was found for the single talker condition and when the number of simultaneous talkers exceeded six. These impressions were supported by post hoc pairwise comparisons (Bonferroni-adjusted *t*-tests).

CONCLUSIONS

The principal conclusion that emerges from the present experiment is that the spatial separation of speech signals in the horizontal plane enhances one's ability to identify critical speech signals when they occur in competing message environments. Specifically, spatial separation significantly increased identification scores when the number of simultaneous talkers was between two and six (see Figure 4) for both the free field and virtual audio conditions. In contrast, detection scores were not mediated by the spatial separation factor. Collectively, these results have important implications for the use of spatialized speech interfaces, especially in application domains in which operators are required to accurately monitor and identify speech signals in competing message environments.





Figures 3 (left) and 4 (right). Mean percent correct detections for the male and female critical call sign across the eight talker conditions. Mean percent correct identifications for spatialized and non-spatialized conditions across the eight talker conditions.

It is also interesting to point out that these data imply a potential means for initiating and executing adaptive interfaces for multi-channel communications. Over the past decade, researchers (Hancock & Chignell, 1988; Hettinger, Cress, Brickman, & Haas, 1996) have speculated on the effectiveness of adapting display and control interfaces to augment operator performance, particularly in complex task environments which challenge, or exceed, the perceptual, perceptual-motor, and/or cognitive capacities of the operator. In brief, the goal of adaptive interface design is to be able to modify dynamically the display and/or control characteristics of a human-machine interface in response to changes in the functional state of the operator, the machine, and/or the external environment. Recent examples of adaptive interface design for airborne applications include multisensory navigation aides (Moroney, 1999) and unisensory and multisensory target acquisition displays (Tannen, 1999). The data reported herein suggest several candidate adaptation schemes for communication interfaces.

Number of Talkers. Clearly, the increase in intelligibility due to spatialization is a function of the number of talkers. As illustrated in Figure 4, enhanced performance on the identification task only occurs when there are between two and six talkers speaking simultaneously. Hence a meaningful adaptation might involve spatializing speech signals only when the number of competing messages is within this range.

Sex of Talker. While detection performance was found to vary inversely with the number of talkers, there was a significant advantage for the detection of critical call signs spoken by female talkers. Although this effect was not mediated by spatialization, it does suggest a potential interface adaptation that may be particularly effective for the presentation of synthesized speech warnings. Specifically, the saliency and detectability of warnings may be enhanced by employing a female-voiced warning signal under the appropriate masking conditions (see Ericson and McKinley, 1997, for review of sex of talker effects in speech masking).

Spatial Location of Competing Audio Signals. Earlier research (Nelson, Bolia, Ericson, & McKinley, 1998a; 1998b) focusing on the spatial location and distribution of free field and virtual speech signals indicated only an additive benefit of spatialization – that is, enhanced intelligibility did not depend on either the location or the distribution of the competing speech signals. Accordingly, as suggested by Bolia and Nelson (in press) there may be cases in which it is advantageous to shift the spatial location of speech signals away from additional auditory and/or visual information co-located therewith, in order to preserve the integrity of both the communications and the competing information.

Introduction of Temporal Asynchrony. One of the limitations of the methodology employed in the present investigation centers around the fact that the competing speech signals were always presented simultaneously. While this was useful for isolating the effects of spatialization, it seems intuitive that the introduction of temporal asynchrony between the onsets of the phrases may further enhance these effects. Indeed, recent research has shown this to be the case (Simpson, Ericson, Bolia, & McKinley, 1999). As such, it may be useful to adaptively time-shift the onsets of competing speech signals when the number of talkers exceeds a certain criterion.

References

Begault, D. R., & Pittman, M. T. (1996). Three-dimensional audio versus head-down traffic alert and collision avoidance system displays. *The International Journal of Aviation Psychology*, *6*, 79-93.

Begault, D. R., & Wenzel, E. M. (1993). Headphone localization of speech. *Human Factors*, *35*, 361-376.

Bolia, R. S. & Nelson, W. T. (in press). Spatial audio displays for target acquisition and speech communication. In L. J. Hettinger, & M. W. Haas (Eds.), *Psychological Issues in the Design and Use of Virtual Environments*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Crispien, K., & Ehrenberg, T. (1995). Evaluation of the "cocktail party effect" for multiple speech stimuli within a spatial auditory display. *Journal of the Audio Engineering Society*, *43*, 932-941.

Ericson, M. A., & McKinley, R. L. (1997). The intelligibility of multiple talkers separated spatially in noise. In R. H. Gilkey, & T. R. Anderson (Eds.), *Binaural and spatial hearing in real and virtual environments*. (pp. 701-724). Mahwah, NJ: Lawrence Erlbaum Associates.

Hancock, P. A., & Chignell, M. H. (1988). Mental workload dynamics in adaptive interface design. *IEEE Transactions on Systems, Man, and Cybernetics, 18(4)*, 647-657.

Hettinger, L. J., Cress, J. D., Brickman, B. J., & Haas, M. W. (1996). Adaptive interfaces for advanced airborne crew stations. *Proceedings of the Third Annual Symposium on Human Interaction with Complex Systems* (pp. 188-192). Los Alamitos, CA: IEEE Computer Society Press.

McKinley, R. L., Ericson, M. A., D'Angelo, W. R. (1994). 3-dimensional auditory displays: Development, applications, and performance. *Aviation, Space, and Environmental Medicine*, *May*, 31-38.

Moore, T. J. (1981). Voice communication jamming research. *AGARD Conference Proceedings 311: Aural Communication in Aviation* (pp. 2:1-2:6). Neuilly-Sur-Seine, France.

Moroney, B. W. (1999). An evaluation of unisensory and multisensory adaptive flight path navigation displays. Unpublished doctoral dissertation. University of Cincinnati, Cincinnati, Ohio.

Nelson, W. T., Bolia, R. S., Ericson, M. A., & McKinley, R. L. (1998a). Monitoring the simultaneous presentation of multiple spatialized speech signals in the free filed. *Proceedings of the 16th International Congress on Acoustics and the 135th Meeting of the Acoustical Society of America* (pp. 2341-2342). Woodbury, NY: Acoustical Society of America.

Nelson, W. T., Bolia, R. S., Ericson, M. A., & McKinley, R. L. (1998b). Monitoring the simultaneous presentation of spatialized speech signals in a virtual acoustic environment. *Proceedings of the 1998 IMAGE Conference* (pp. 159-166). Chandler, AZ: The IMAGE Society, Inc.

Perrott, D. R., Cisneros, J., McKinley, R. L., & D'Angelo, W. R. (1996). Aurally aided visual search under virtual and free field listening conditions. *Human Factors*, *38*, 702-715.

Ricard, G. L., & Meirs, S. L. (1994). Intelligibility and localization of speech from virtual directions. *Human Factors*, *36*, 120-128.

Simpson, B. D., Bolia, R. S., Ericson, M. A., & McKinley, R. L. (1999). The effect of sentence onset asynchrony on call sign detection and message intelligibility in a simulated "cocktail party." *Journal of the Acoustical Society of America*, *105*, 1024.

Tannen, R. S. (1999). Adaptive integration of helmet-mounted and spatialized auditory displays for target localization. Unpublished doctoral dissertation. University of Cincinnati, Cincinnati, Ohio.

Wenzel, E. M. (1992). Localization in virtual acoustic displays. *Presence*, *1*, 80-106.

Yost, W. A., Dye, R. H., & Sheft, S. (1996). A simulated "cocktail party" with up to three sound sources. *Perception & Psychophysics*, *58*, 1026-1036.